Generalized competitive clustering for image segmentation

Boujemaa, N.

INRIA, Rocquencourt;

This paper appears in: **Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American**

Meeting Date: 07/13/2000 -07/15/2000

Publication Date: 2000

Location: Atlanta, GA , USA

On page(s): 133-137

References Cited: 13

IEEE Catalog Number: PR00789

Number of Pages: xi+496

INSPEC Accession Number: 6748767

Abstract:

We focus on the problem of unsupervised clustering which allows automatic setting of the optimal cluster number. We present a generalization of the competitive agglomeration clustering algorithm firstly introduced in (Frigui and Krishnapuram, 1997). This generalization is inspired by the regularization theory and suggests a new schema for using various cluster validity criteria continuously proposed in the literature. As a consequence of this generalization, we introduce new objective clustering functions, and present their associated optimal solutions. We present an application of this competitive clustering schema to color image segmentation in order to perform partial queries in the context of image retrieval by content. In this case, each pixel is represented by the color distribution in its vicinity. The Clustering algorithm has to incorporate an appropriate distance measure to compare feature vector similarity

Index Terms:

competitive algorithms content-based retrieval image colour analysis image segmentation visual databases cluster validity criteria color image segmentation competitive agglomeration clustering algorithm content based image retrieval feature vector similarity generalized competitive clustering image segmentation objective clustering functions pixel regularization theory unsupervised clustering

Documents that cite this document

Select link to view other documents in the database that cite this one.

Generalized competitive clustering for image segmentation

Nozha Boujemaa
INRIA Rocquencourt
Nozha.Boujemaa@inria.fr

Abstract

In this paper, we focus on the problem of unsupervised clustering which allows automatic setting of optimal clusters number. We present a generalization of the competitive agglomeration clustering algorithm firstly introduced in [1]. This generalization is inspired by the regularization theory and suggests a new schema for using various cluster validity criteria continuously proposed in the literature. As a consequence of this generalization, we introduce new objective clustering functions, and present their associated optimal solutions.

We present an application of this competitive clustering schema to color image segmentation in order to perform partial queries in the context of image retrieval by content. In this case, each pixel is represented by the color distribution in its vicinity. Clustering algorithm has to incorporate an appropriate distance measure to compare feature vectors similarity.

1. Introduction

Our work was motivated by requirements and constraints in the context of image retrieval by content. Most systems use the query-by-example approach, performing queries such as "show me more images that look like this one". Most often, the user is more specifically interested in specifying an object (or region) and in retrieving more images with similar objects (or regions), as opposed to similar images as a whole. Our aim is to allow the user to perform a query on image parts (region of interest). Methods range from manual region delimitation [11], systematic image subdivision without segmentation [10] to approximate region segmentation [9]. In this paper, we address the problem of clustering based segmentation of each image in the database to allow partial queries.

Most popular clustering algorithms have the drawback to need a predefined number of clusters. Since image databases are often huge, the prior setting of clusters number for each image is no longer viable. This context requirement motivates our interest to such idea of unsupervised clustering. In the first section, we describe existing competitive clustering algorithm. We address the issue of explanation and generalization in the second section. Update equations computation and interpretation are given in section 3. We deal with color image description and segmentation and show some results in section 4.

2. Initial algorithm

Competitive agglomeration clustering (CA) schema was firstly proposed in [1] to minimize the following objective function

$$J=\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^2(d_{ij})^2-\alpha\sum_{i=1}^{C}\left[\sum_{j=1}^{N}\mu_{ij}\right]^2$$

with the constraint:

$$\sum_{i=1}^{C}\mu_{ij}=1 \quad \forall\, 1\le j\le N \text{ and } \mu_{ij}\in[0,1]$$

where (μ_{ij}) is the membership degree of the j-th data point x_j in the i-th cluster, d_{ij} is their distance, N is the total number of gray levels and C the number of clusters to be found (which will be dynamically updated). The initial CA partition has an over-specified number of clusters, which is dynamically reduced as the algorithm progresses. At the convergence, the final partition has the optimal number of clusters.

The objective function combines two components. The first one is similar to the FCM [3] objective function and has a global minimum when each data point is in a separate cluster. The global minimum of the second component is achieved when all points are in the same cluster such that it controls the number of clusters. The two components are combined by the parameter α which is chosen to equilibrate their contributions:

$$\alpha(k)=\eta_0 . e^{-k/\tau} . \frac{\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^2.(d_{ij})^2}{\sum_{i=1}^{C}\left[\sum_{j=1}^{N}\mu_{ij}\right]^2}$$

The value of α decreases slowly such that it favors agglomeration in the first iterations while emphasizing objective function in the latest ones.

The update equation of memberships is :

$$\mu_{ij} = \frac{1/d_{ij}^2}{\sum_{k=1}^{C} 1/d_{kj}^2} + \frac{\alpha\left(N_i - \tilde{N}_j\right)}{d_{ij}^2}$$

and can be written as:

$$\mu_{ij} = \mu_{ij}^{FCM} + \mu_{ij}^{Bias}$$

were:

$$N_i = \sum_{j=1}^{N} u_{ij} \text{ and } \tilde{N}_j = \sum_{k=1}^{C} \frac{1}{d_{kj}^2} N_k \Big/ \sum_{k=1}^{C} \frac{1}{d_{kj}^2},$$

are respectively the cardinality of cluster i and a weighted average of the cluster cardinality where the weight reflects its proximity to the data point x_j.

μ_{ij}^{FCM} is the membership of the objective function FCM [3] while the sign of μ_{ij}^{Bias} express the competition between clusters and leads to gradual reduction of the cardinality of spurious clusters. A given cluster vanishes when its cardinality N_j is less than a minimum required.

On the other hand, another similar unsupervised clustering algorithm was proposed *separately* in the bayesian framework based on entropy minimization[2]:

$$U = \sum_{i=1}^{k} \sum_{j \in S_i} \left[\frac{(x_j - y_i)^2}{2\sigma^2} - \frac{\alpha_E}{N} \ln p_i \right]$$

When $\alpha_E = 0$, U is the energy of k-means algorithm. We aim to explain the relationship between both algorithms in the following section.

3. Generalization

Both objective functions aim to minimize a J criterion that we write as:

$$J = J_1 + \alpha J_2$$

J is the combination of two terms of *antagonist behavior:*

- J_1 express the fidelity to the data. This term is minimum when each data point constitute a separate cluster,
- J_2 is a complexity reduction term. It is minimal when all data points are in the same cluster. Uncertainty is maximal in this case.

Note that this schema is similar to energy minimization model [13][12] when approximation function is piece wise constant.

Optimal cluster number is determined by a balance between these two opposite effects terms. We note that J2 reduce complexity of data partition in the feature space. We would like here to emphasize that cluster validity theory has exactly the same goal. Several criteria are continuously proposed in the literature to estimate the quality of a given partition by reducing partition complexity. Also, we note that the second term used in the bayesian framework, consisting of an entropy measure, has been already proposed by Bezdek [3] as a validity criterion. It is clear now that all validity criteria are possible instantiations of J2 term. This allow us to explain the relationship between both initial algorithms and suggest a family of possible other J2 terms. Remember that, in the literature, validity criteria are used sequentially is as follows:

- compute different data partition for c= 2, ... cmax
- at each convergence, compute the value of validity criteria
- seek for the extremum value of validity criterion and set the optimal number of cluster to its correspondent c value.

With the new proposed schema, validity criterion are taken into account "in parallel" i.e. at the same time with partition allowing only one convergence process computation.

In the following section we dress the solution for two new objective function based on entropy measures. We notice that they are not unique and they have the advantage to be not numerically complex.

4. Optimization and interpretation

We present optimization procedure with two validity criteria proposed yet in [3]. They refer to information theory and have the advantage to be less complex, for optimization, than many validity criteria more recently proposed [5] [6]. Entropy measures have a maximal values when partition is completely "fuzzy" and membership matrix is constant and identical to 1/c.

$$J = \sum_{i}^{c} \sum_{j}^{N} \mu_{ij}^2 d_{ij}^2 - \frac{\alpha}{N} \sum_{i}^{c} \sum_{j}^{N} \mu_{ij} \log\left(\mu_{ij}\right)$$

$$\text{with} \quad \sum_{i=1}^{c} \mu_{ij} = 1 \quad \forall j$$

we obtain, λ_j are Lagrange multipliers:

$$L(J) = J + \sum_{j}^{N} \lambda_j \left(\sum_{i}^{c} \mu_{ij} - 1 \right)$$

$$\frac{\partial Lj}{\partial \mu_{ij}} = 2\mu_{ij} d_{ij}^2 - \frac{\alpha}{N}\left(1 + \log \mu_{ij}\right) + \lambda_j = 0.$$

$$\mu_{ij} = \frac{1}{2d_{ij}^2}\left[-\lambda_i + \frac{\alpha}{N}\left(1 + \log \mu_{ij}\right)\right]$$

$$\sum_{i}^{c} \mu_{ij} = 1 = \sum_{i=1}^{c}\left(\frac{-\lambda_j}{2d_{ij}^2}\right) + \frac{\alpha}{2N}\sum_{i=1}^{c} \frac{1 + \log \mu_{ij}}{d_{ij}^2}$$

Solving this equation for λj leads to:

$$-\lambda_j = \left[1 - \frac{\alpha}{2N}\sum_{s=1}^{c}\left(\frac{1 + \log \mu_{sj}}{d_{sj}^2}\right)\right]\frac{2}{\sum_{s}^{c} \frac{1}{d_{sj}^2}}$$

Substituting λj in the expression of uij, we obtain the following update equation for the membership of a data point xj to cluster j:

$$\mu_{ij} = \frac{1/d_{ij}^2}{\sum_{s}^{c}\left[\frac{1}{d_{sj}^2}\right]} + \frac{1}{d_{ij}^2}\frac{\alpha}{2N}\left[\log \mu_{ij} - \frac{\sum_{s}^{c}\left(\frac{\log \mu_{sj}}{d_{sj}^2}\right)}{\sum_{s}^{c} 1/d_{sj}^2}\right]$$

As the initial algorithm, we obtain:

$$\mu_{ij} = \mu_{ij}^{Fcm} + \mu_{ij}^{Biais}$$

$$\text{where: } \mu_{ij}^{Fcm} = \frac{1/d_{ij}^2}{\sum_{s}^{c}\left[\frac{1}{d_{sj}^2}\right]}$$

and:

$$\mu_{ij}^{Biais} = \frac{1}{d_{ij}^2}\frac{\alpha}{2N}\left[\log \mu_{ij} - \left(\overline{\log \mu_{sj}}^{s}\right)_j\right]$$

$$\text{with: } \left(\overline{\log \mu_{sj}}^{s}\right)_j = \frac{\sum_{s}^{c}\left(\frac{\log \mu_{sj}}{d_{sj}^2}\right)}{\sum_{s}^{c} 1/d_{sj}^2}$$

that represents the weighted average of $\log \mu_{sj}$ over all clusters s.

The sign of μ_{ij}^{Biais} allows the competition process, as in the initial algorithm, by reinforcement or reduction of cardinality cluster.

In the same way, we present update equation for statistical entropy.

$$J = \sum_{i}^{c}\sum_{j}^{N} \mu_{ij}^2 d_{ij}^2 - \alpha \sum_{i}^{c} p_i \log p_i$$

with:

$$\sum_{i=1}^{c} \mu_{ij} = 1 \quad \forall j \quad \text{and} \quad p_i = \frac{1}{N}\sum_{j=1}^{N} \mu_{ij} = \frac{N_i}{N}$$

Ni is the cardinality of cluster i. In the same way as the previous optimization procedure for entropy partition, we finally obtain the following membership updating equation:

$$\mu_{ij} = \frac{1/d_{ij}^2}{\sum_{s}^{c}\left(1/d_{sj}^2\right)} + \frac{\alpha}{2N d_{ij}^2}\left[\log p_i - \frac{\sum_{s}^{c} \frac{1}{d_{sj}^2}\log p_s}{\sum_{s}^{c} 1/d_{js}^2}\right]$$

where $p_i = \frac{N_i}{N}$ and $p_s = \frac{N_s}{N}$ represent resp. probabilty of cluster i and s or relative cardinalties of these clusters.

$$\mu_{ij} = \frac{1/d_{ij}^2}{\sum_{s}^{c}\left(1/d_{sj}^2\right)} + \frac{\alpha}{2N d_{ij}^2}\left[\log p_i - (\overline{\log p_s}^{s})_j\right]$$

$$\text{with:} \quad (\overline{\log p_s}^{s})_j = \overline{P}_j = \frac{\sum_{s}^{c} \frac{1}{d_{sj}^2}\log p_s}{\sum_{s}^{c} 1/d_{js}^2}$$

Partition entropy has the advantage to be more robust to parameter α initialization than the statistical entropy.
Note that, in both cases, we have the same effect of competition described by the sign of μ_{ij}^{Biais} given by a difference between the current cluster membership parameter and a particular weighted average of the other clusters memberships where the weight reflects its proximity to the current data point.

5. Application to image segmentation

We apply the CA schema to color image segmentation in the context of content based image retrieval. Segmentation allows performing a query on a region of interest instead of the usual query on the whole image (query by example). Indeed, in this context we need an unsupervised clustering algorithm since we could not set the number of clusters for each image. For each pixel, we consider color distribution represented by its local 3D histogram as color feature. Then the distance measure for clustering is similar to the distance used in [6] and is given by:

$$d_{L_1}(x,y) = \sum_{t=1}^{b} |h_t(x) - h_t(y)|$$

where b is the number of histogram bins and $h_t(x)$ is the local 3D color histogram in the vicinity of the pixel x.
We use gradual focusing decision [4] after partition convergence. In the following pictures, we present typical segmentation results.

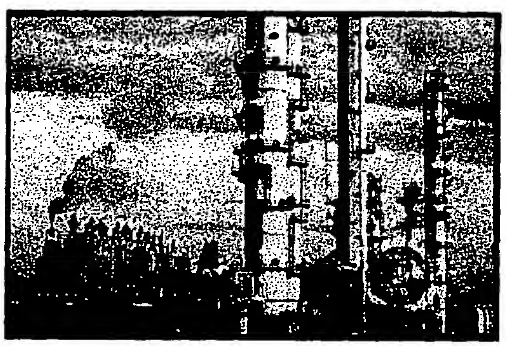

fig. 1 Original color image

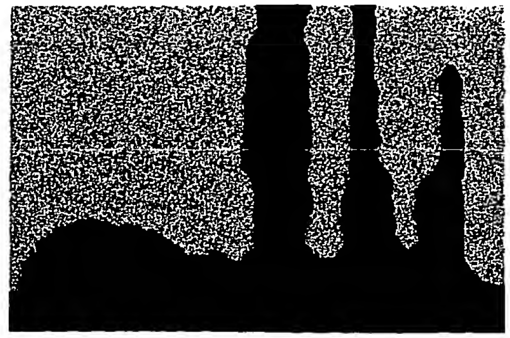

fig. 2 Segmentation result by local 3D color histogram competitive clustering

More segmentation results are shown on http://www-rocq.inria.fr/~boujemaa/Partielle2.html . The obtained regions will be used as mask query on which Surfimage [8] image signatures will be computed to seek for partial queries similarity.

fig. 3 Segmentation result using only color components as feature vector

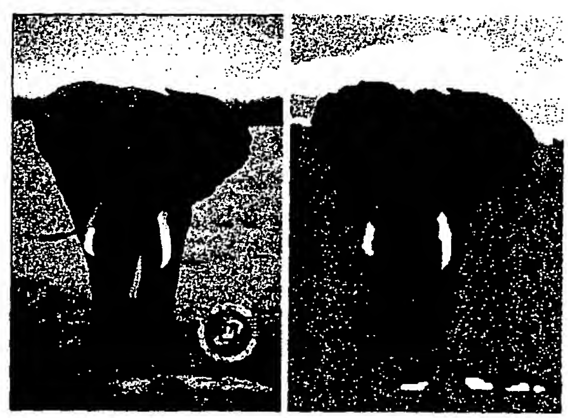

fig. 4 Original images and its clustering based segmentation

6. Conclusion

In this paper, we have address the generalization of competitive agglomeration clustering algorithm[1] that explains the relationship with another bayesian clustering algorithm [2]. This generalization suggests a new use of cluster validity criteria continuously presented in the literature. As a consequence, we propose new objective functions that are not unique. We present results for color image segmentation in the context of image retrieval by content.

7. References

[1] Frigui H., Krishnapuram R., "Clustering by Competitive Agglomeration" , Pattern Recognition Letters, 07-1997, 30 (7), 1109-1119

[2] Palubinskas G, Descombes X, Kruggel F, "An unsupervised Clustering Method using the entropy Minimization", CVPR'98, pp 1816 – 1818

[3] Bezdek J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981

[4] N. Boujemaa and G. Stamon, "Fuzzy Modeling in Early Vision - Application to Medical Image Segmentation" in Progress in Image Analysis and Processing III, pp: 649-656. Ed. S. Impedovo - World Scientific 1994

[5] Xie X and Beni G., "A validity measure for fuzzy clustering", IEEE Trans. on PAMI, 13 (8) : 841-847, 1991

[6] Bezdek J. and Pal N., "Some new indexes of cluster validity", IEEE Trans. on Syst., Man and Cybernetics, Part B, vol. 28, 1998, pp 301-315

[7] Michaeal J. Swain and D. H. Ballard: Color indexing, IJCV,vol.7, n°1, pp.11-32, 1991

[8] Nastar C., Mitschke M., Meilhac C. and Boujemaa N.: "Surfimage: a Flexible Content-Based Image Retrieval System", ACM Multimedia'98, Bristol, Sept. 1998

[9] Belongie S. and Carson C. and Greenspan H. and Malik J.: Color- and Texture-Based Image Segmentation Using EM and Its Application to Content-Based Image Retrieval", Proceedings of the Sixth International Conference on Computer Vision(ICCV'98), Bombay, January,1998

[10]J Malki, N. Boujemaa, C. Nastar and A. Winter, "Region Queries without Segmentation for Image Retrieval by Content", Third International Conference on Visual Information Systems (Visual'99), Amsterdam, The Netherlands, June 2-4, 1999, Lecture Notes in Computer Science vol. 1614, p 115-122

[11]Del Bimbo and Vicario E. : Using Weighted Spatial Relationships in Retrieval by Visual Contents, IEEE workshop on Image and Video Libraries",Santa Barbara,June,1998

[12]Geman S., Geman D., « Stochastic relaxation, Gibbs distribution and the bayesian restauration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6 (6) : 721-741, 1984

[13]D. Mumford & J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems", Communications on Pure and Applied Mathematics. vol. XLII No.4, 1989.

Proceedings of the 1991 IEEE International Symposium
on Intelligent Control
13 - 15 August 1991, Arlington, Virginia, U.S.A.

A NEW FUZZY CLUSTERING VALIDITY CRITERION AND ITS APPLICATION TO COLOR IMAGE SEGMENTATION

Xuanli lisa Xie and Gerardo Beni
Center for Robotic Systems
University of California, Santa Barbara, CA 93106

I. Introduction

The engineering literature has paid very little attention to cluster validity issues [1], limiting the effort to present new clustering algorithms which perform reasonably well on a few data sets. In particular, the issue of validity for clustering of fuzzy data sets has been neglected (with few notable exceptions [2,3]). On the other hand, if fuzzy cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to fundamental questions of cluster tendency. Recently, validity of fuzzy clustering has been discussed in applications to mixtures of normal distributions [4,5]. Also applications to distributed perception [6] have been proposed which rely in an essential way on good validity criteria for fuzzy clustering.

In the latter applications [7], separated sensors observe a common object. They communicate to a central processor not (perceptual) data (which, due to their size, cannot be transmitted in real time) but decisions (which, due to their smaller bit-size, can be transmitted in real time). In such cases, a fundamental decision is often the determination of the number of 'objects' observed, i.e. the validity of the clustering procedure. Since higher level decisions by the central processor are based on the validity of these separated clustering procedures, it is essential that an efficient method is developed for fuzzy clustering validity.

Generally the issue of cluster validity is a broad one and involves many questions. In view of the applications to distributed perception, in this paper we focus on the validity of a partition. The answer is sought, as is generally accepted [1,3], in measures of separation among clusters and cohesion within clusters.

II. Clustering Algorithm and Validity Criteria

Clustering is a tool that attempts to assess the relationships among patterns of the data set by organizing the patterns into groups or clusters such that patterns within a cluster are more similar to each other than are patterns belonging to different clusters.

There are many clustering algorithms [2,3,8], but an important question in clustering is 'cluster validity' which deals with the significance of the structure imposed by a clustering method. Performance of many existing clustering algorithms are studied in [9]. We will briefly review Fuzzy c-means clustering algorithm below for later reference.

A. Fuzzy c-means clustering algorithms

The fuzzy c-means (FCM) clustering algorithm (Bezdek [2]) is the fuzzy equivalent of the nearest mean "hard" clustering algorithm (Duda and Hart [10]), minimizes the following objective function with respect to fuzzy membership μ_{ij} and cluster centroid V_i.

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m d^2(X_j, V_i), \tag{1}$$

where

$$d^2(X_j,V_i) = (X_j-V_i)^T A (X_j-V_i) \tag{2}$$

A is a p×p positive definite matrix, p is the dimension of the vectors X_j ($j = 1,2,\ldots,n$), c is the number of clusters, n is the number of vectors (or data points), m>1 is the fuzziness index [2].

The FCM algorithm is executed in the following steps [2]:

1. Initialize memberships μ_{ij} of X_j belonging to cluster i such that

$$\sum_{i=1}^{c} \mu_{ij} = 1 \tag{3}$$

2. Compute the fuzzy centroid V_i for $i = 1,2,\ldots, c$ using

$$V_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m X_j}{\sum_{j=1}^{n} (\mu_{ij})^m} \tag{4}$$

3. Update the fuzzy membership μ_{ij} using

$$\mu_{ij} = \frac{\left(\frac{1}{d^2(X_j,V_i)}\right)^{\frac{1}{(m-1)}}}{\sum_{i=1}^{c} \left(\frac{1}{d^2(X_j,V_i)}\right)^{\frac{1}{(m-1)}}} \tag{5}$$

4. Repeat steps (3) and (4) until the value of J_m is no longer decreasing.

The FCM algorithm always converges to strict local minima of J_m starting from an initial guess of μ_{ij}, but different choices of initial μ_{ij} might lead to different local minima [4]. This algorithm is more suitable for clustering according to a geometric distance measure, while other algorithm, e.g. Maximum-Likelihood Estimation algorithm [10,11] may be more suitable in other cases

B. Validity criteria for hard and fuzzy clustering

A well established hard cluster validity criterion is the separation indices D_1 (Dunn [13]) which identifies 'compact, separate' (CS) clusters and is defined by

$$D_1 = \min_{i+1\le j\le c-1} \left\{ \min_{1\le i\le c} \left(\frac{dis(u_i, u_j)}{\max_{1\le k\le c} \{ dia(u_k) \}} \right) \right\} \tag{6}$$

where

$$dia(u_k) = \max_{X_i,X_j \in u_k} d(X_i,X_j) \tag{7}$$

$$dis(u_i,u_j) = \min_{X_i\in u_i, X_j\in u_j} d(X_i,X_j) \tag{8}$$

d is any metric induced by an inner product on R^p. The validity measure of CS clustering of X solves $\max_{2\le c\le n} \{\max_{\Omega_c} D_1\}$, where Ω_c denotes the optimality candidates at fixed c. It is proved that a hard c-partition of X contains c compact, separate (CS) clusters if $D_1>1$. Furthermore, there is at most one CS partition of X if $D_1>1$. The main drawback with direct implementation of this validity measure is computational since calculating D_1 becomes computationally very expensive as c and n increase. Another validity criterion which also measures compact and separate clusters is introduced by Davies and Bouldin [14]. The major difference from D1 is that it considers the average case by using the average error of each class. Jain and Moreau [15] also defined a method for cluster validity by using a bootstrap technique, that could be used with any clustering algorithm.

As a fuzzy clustering validity function Bezdek [16] designed the partition coefficient F to measure the amount of "overlap" between clusters.

$$F = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^2 \tag{9}$$

In this form F is inversely proportional to the overall average overlap between pairs of fuzzy subsets. In particular, there is no membership sharing between any pairs of fuzzy clusters if F=1. Solving

$\max_c \{ \max_{\Omega_c} \{F\} \}$ (c = 2, 3 ... n-1) is assumed to produce valid clustering of the data set X. Disadvantages of the partition coefficient are the lack of direct connection to a geometrical property and its monotonic decreasing tendency with c. There are several other criteria in the literature which also measure the amount of fuzziness, such as classification entropy [17], proportion exponent [18], uniform data functional [19], non fuzziness index [20] and information ratio [21]. Those criteria share a similar drawback with F, that is the lack of direct connection to the geometrical property of data set.

Gunderson [22] introduced a separation coefficient which takes into account geometrical properties. This validity criterion is designed to identify compact and separated clusters (which is similar to our goal). However, this method cannot be directly applied. It works on fuzzy clustering outputs by first converting them to hard ones. Since there are many ways one can convert fuzzy partitions to hard ones, this method shares the shortcomings of non-uniqueness of transferring from fuzzy partitions to hard partitions.

III. A Compact and Separated Fuzzy Validity Criterion

In this section we define S as a fuzzy clustering validity function which measures the overall average compactness and separation of a fuzzy c-partition. We also give a heuristic rationale and an implementation strategy for the use of this function.

A. Definition of a new fuzzy clustering validity function S

Consider a fuzzy c-partition of the data set $X = \{ X_j; j = 1, 2, \ldots, n\}$ with V_i ($i = 1, 2, \ldots, c$) the centroid of each cluster and μ_{ij} ($i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, n$) as the fuzzy membership of data point j (also called vector j) belonging to class i.

Definition 1: $d_{ij} = \mu_{ij}\|X_j - V_i\|$, is called the *Fuzzy Distance* of X_j to class i.

Note that $\|\cdot\|$ is the usual Euclidean norm. Thus d_{ij} is just the Euclidean distance between X_j and V_i weighted by the fuzzy membership of data point j belonging to class i.

Definition 2: $n_i = \sum_j \mu_{ij}$ is called the *Fuzzy Number* of vectors in class i.

Note that $\sum_i n_i = n$, where n is a 'hard' number, e.g. the total number of data points in X. In the extreme case, when the partition is hard, n_i becomes exactly the number of vectors in class i.

Definition 3: for each class i, the summation of the squares of Fuzzy Distance of each data point, denoted by σ_i, is called the *Variation* of class i, that is: $\sigma_i = \sum_j (d_{ij})^2 = (d_{i1})^2 + (d_{i2})^2 + \ldots + (d_{in})^2$. The summation of the variations of all classes, denoted by σ, is called the *Total Variation* of data set X with respect to the fuzzy c-partition, i.e., $\sigma = \sum_i \sigma_i = \sum_i \sum_j (d_{ij})^2$.

Note that σ_i and σ depend on the data set, but more importantly they depend on the fuzzy c-partition, i.e., μ_{ij}'s and V_i's. A better c-partition should result in smaller σ. These values are not normalized, and they depend on how we choose our coordinate system. For example, if the fuzzy c-partition is obtained by using the fuzzy c-means algorithm with m = 2, the value of σ will be equal to the c-means objective function J_2 (Sec. II.A.1).

Definition 4: the ratio, denoted by π, of the total variation to the size of the data set, that is, $\pi = (\sigma/n)$, is called the *Compactness* of the fuzzy c-partition of the data set.

The value π measures how compact each and every class is. The more compact the classes are, the smaller π is. π is a function of the distribution characteristics of the data set itself, and more importantly a function of how we divide the data points into clusters. But it is independent of the number of data points. For a given data set, a smaller π indicates that we have reached a partition with more compact clusters, thus indicating a better partition. Gath and Geva [4] introduced fuzzy hypervolume which is the probability weighted total variation. This validity measure can identify ellipsoidal clusters and overlapped clusters. By incorporating covariance into the distance matrix A in equation (2), π can also identify ellipsoidal clusters.

Definition 5: the quantity $\pi_i = (\sigma_i/n_i)$ is called the *Compactness of class i*.

Since n_i is the number of vectors in class i, σ_i/n_i is the average variation in class i. We have defined the compactness of fuzzy c-partition in terms of total variation and number of vectors. After defining π_i, we have some alternative ways to define the compactness of the fuzzy c-partition, such as: $\pi = (\sum_i \pi_i)/c$, i.e., the average compactness of each class; or $\pi = \max \pi_i$, i.e., the worst case. It can be shown that both ways have similar effect to definition 4.

Definition 6: $s = (d_{min})^2$ is called the *Separation* of the fuzzy c-partition, where d_{min} is the minimum distance between cluster centroids, i.e.:

$$d_{min} = \min_{ij} \|V_i - V_j\|$$

A larger s indicates that all the clusters are separated.

Definition 7: the *Compactness and Separation Validity Function* S is defined as the ratio of compactness π to the separation s, i.e. $S = \pi/s$.

After substituting for π and s, we get $S = (\sigma/n) / (d_{min})^2$. A smaller S indicates a partition in which all the clusters are overall compact, and separate with each other. Thus, our goal is to find the fuzzy c-partition with the smallest S.

S can be explicitly written as:

$$S = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^2 \|V_i - X_j\|^2}{n \min_{ij} \|V_i - V_j\|^2} \tag{10}$$

We note that the definition of S is independent of the algorithm used to obtain μ_{ij}. Thus it is not internal to the clustering algorithm. For FCM algorithm with m=2, S can be shown to be:

$$S = \frac{J_2}{n \ (d_{min})^2} \tag{11}$$

which is very easy to calculate. More importantly, minimizing S corresponds to minimizing J_2, which is the goal of FCM. The additional factor in S is $(d_{min})^2$, which is the separation measurement. The more separate the clusters, the larger $(d_{min})^2$, and the smaller S. Thus, the smallest S indeed indicates a valid optimal partition.

We note, however, that S is still monotonically decreasing when c gets very large and close to n (under certain statistical assumptions). One thing we can do is to impose an *ad hoc* punishing function [23] to eliminate this decreasing tendency. How to choose this function is not discussed here. Nevertheless, we shall see that even without a punishing function the validity function S provides a well defined method to solve the validity problem.

There are some existing validity criteria in the literature which measure compact and separate clustering. The separation coefficient in [22] considers the worst case while S is more interested in total average case. Furthermore, the separation coefficient cannot be directly applied to fuzzy clustering as mentioned before. In [14], Davies and Bouldin introduces a hard partition validity criterion R. It is roughly related to S by R=S/c if S is used for hard partitions. However, from our experience, S/c has a decreasing tendency as c increase.

B. Case of validity function S < 1/8

As discussed previously, an identifiable substructure results in a small S, but it is not immediately obvious how small S is. On the other hand, it is possible to find a heuristic threshold S_0 such that if $S < S_0$ the fuzzy partition is overall compact and separate.

Consider two circular clusters with radius R and uniform vector density ρ. If the c-partition is separated, then

$$R < \frac{1}{2} d_{min} \tag{12}$$

where d_{min} is as above. Let ΔA be a small area inside a cluster and r be the distance from ΔA to V_i, the cluster centroid. Thus the variation σ_i of the cluster i can be written as:

$$\sigma_i = \int\int r^2 \rho \Delta A$$

The number of vectors in the cluster i is $n_i = \pi R^2 \rho$. Let us suppose that the data set is composed of classes of equal variance, then the total variation σ can be written as $\sigma = c\sigma_i$ and the total number of vectors $n = cn_i$. Thus the validity function can be written as

$$S = \frac{1}{\pi R^2 \rho * (d_{min})^2} \int\int r^2 \rho \Delta A = \frac{1}{2 * (d_{min})^2} R^2$$

If the clusters are separate, according to equation (12),

$$S < \frac{1}{2 * (d_{min})^2} * \frac{1}{4} (d_{min})^2 = \frac{1}{8}. \tag{13}$$

Thus we may consider the c-partition as compact and separate if S < 1/8. We have assumed that the data is uniformly distributed in obtaining the value S_0=1/8. However, the actual data set will not always be of this kind. Thus we should not attribute too much significance to the actual value S_0=1/8. If we use a more strict density function or choose the density function with pre-knowledge about the data, the value of S_0 will get smaller. This value just gives us a guideline and S < 1/8 is not a necessary nor a sufficient condition. Nevertheless, from this result we are led to believe that there exists a certain value such that S less than this value corresponds to a unique partition and, more importantly, that the fuzzy c-partition for S less than this value is a global optimal partition (but it does not mean that S > 1/8 implies clusters that are not compact and separate). We shall see that this is actually the case in Section IV.A.

C. Minimization of S

Since smaller S means a more compact and separate c-partition, we assume that the minimum S partition is the most valid. Thus, a heuristic strategy to use S as a validity function is as follows. Using any fuzzy clustering algorithm, find one or more optimal c-partitions of the data set X for each c = 2, 3, ..., n-1. Let Ω_c denote the optimality candidates at each c, then the solution of $\min_{2\le c\le n-1} \{ \min_{\Omega_c} S \}$ is assumed to yield the most valid fuzzy clustering of the data set X.

D. Implementation strategy

Once we have defined the validity function S, our implementing strategy can be summarized into the following pseudo-algorithm:

1. initialize c <-- 2, S* <-- ∞, c* <-- 1;
2. initialize fuzzy membership μ_{ij};
3. use any stable fuzzy clustering algorithm to update centroids V_i and μ_{ij};
4. do convergence test; if negative goto 3;
5. compute function S;
6. if S < S*, S* <-- S, c* <-- c;
7. if optimal candidate not found, goto 2;
8. c <-- c+1, if c= stop-value, stop;
9. goto 2;

Steps 2-4 are the fuzzy c-partition algorithm. For FCM, equations (4) and (5) can be used. The convergence test can be $(J_m)_{q+1} - (J_m)_q <$ epsilon (e.g. 0.001), where q is an iteration index and J_m is as in equation (1). With m = 2, the S can be easily calculated as in equation (11).

In step 2, the initial values of μ_{ij} can be assigned randomly and then normalized to satisfy $\sum_i \mu_{ij} = 1$ for all j. Another way to initialize μ_{ij} for c > 2 is as follows. When the c partition is increased to c+1, the guess of initial membership for the current c+1 partition can be obtained by assigning the fuzziest vector in previous c partition to be the most determinant vector in the current partition. This method can be described as follows:

$\mu_{c+1\,j} = 0$ for $j \neq k$;
$\mu_{c+1\,k} = 1$;
$\mu_{i\,k} = 0$ for i = 1, ..., c;

where

$k = \arg\min_j \{\mu_{dj}\}$ j = 1,..., n.
$\mu_{dj} = \max_i \{\mu_{ij}\} - \min_i \{\mu_{ij}\}$.

Note that this second method is just another heuristic way for initialization. In many cases, it proves to lead to faster convergence. We may try to mix these two methods together to take advantage of their merits.

A problem of implementation is that S will have a tendency to eventually decrease when c is large. So the value of S is meaningless when c gets close to n. Actually, this is not a serious problem since this phenomenon will not appear in a quite large range of c and since the number of clusters in the clustering problem usually is much smaller than the number of data points. Thus we can use the following three heuristic methods to determine the stop-value of step 8.

First as mentioned in Section III.A we can use a punishing function which imposes on S to counter this decreasing tendency. In Dunn [23], the 'normalization and standardization of a validity function' is a simple example of the idea of punishing function. The second method is that of plotting the optimal value of S for c =2 to n-1, then selecting the starting point of monotonically decreasing tendency as the maximum c to be considered. Let c_{max} denote such a c, then we find c by solving $\min_{2\le c\le c_{max}} \{ \min_{\Omega_c} S \}$. The third way is application dependent. For most applications we do not need to compute S for very large c. It is almost always the case that c at the stop-value is << n. In this instance, we can either choose the maximum c according to pre-knowledge or e.g. let c_{max} = n/3 which very likely would not reach the starting point of the decreasing tendency.

IV. Mathematical Justifications

We have already defined the new validity function and given an implementation strategy to use this function. In this section, we will mathematically justify this new fuzzy validity function via its relationship to a well established hard partition validity measure, and also give numerical examples with the comparison results in section V.

A. Uniqueness and global optimality of the c-partition

The separation index D_1 (proposed by Dunn [13]) is a hard c-partition clustering validity criterion. If D_1>1, unique compact and separated hard clusters have been found. This result turns out to be useful also for fuzzy clustering validity. In fact we may expect that, if the data set X really has distinct substructure, i.e. hard clusters, a fuzzy partitioning algorithm should produce relatively hard memberships μ_{ij} and small total variations. We can prove that if the optimal solution D_1 becomes sufficiently large, the optimal validity function S will be very small, which means that a unique c-partition has been found. The proof of this is as follows.

Definition 8: Let μ_{ij} (i = 1, ..., c; j = 1, ..., n) be the membership of any fuzzy c-partition. The *corresponding hard c-partition* of μ_{ij} is defined as ω_{ij}:

for j = 1, 2, ..., n:
$\omega_{ij} = 1$ if $i = \arg\max_i \{\mu_{ij}\}$;
$\omega_{ij} = 0$ otherwise.

Theorem 1: For any c = 2, ..., n-1, let S be the overall compact and separated validity function of any fuzzy partition, and D_1 be the separation index of the corresponding hard partition, then we have:

$$S \le \frac{1}{(D_1)^2}.$$

Proof: Let the fuzzy c-partition be an optimal partition of the data set X= { X_j; j = 1,2,...,n } with V_i (i =1,2,...c) the centroids of each class u_i, and μ_{ij} the fuzzy membership of the data points X_j belonging to class u_i. The total variation σ_{opt} of the optimal fuzzy c-partition is defined in definition 3. Thus the total variation σ_h of the corresponding hard c-partition is

$$\sigma_h = \sum_i \sum_{X_j \in u_i} \| X_j - V_i \|^2.$$

From the definitions of σ_{opt} and σ_h above, we can get:

$$\sigma_{opt} \le \sum_i \sum_{X_j \in u_i} \| X_j - V_i \|^2.$$

Suppose that the centroid V_i is inside the boundary of cluster i for i = 1 to c. Then

$$\| X_j - V_i \|^2 \leq \mathrm{dia}^2(u_i)$$

for $X_j \in u_i$, where $\mathrm{dia}(u_i)$ is defined in equation (7). We thus have

$$\sigma_{opt} \leq \sum_i \sum_{X_j \in u_i} \mathrm{dia}^2(u_i)$$
$$\leq \sum_i n_i \,\mathrm{dia}^2(u_i)$$
$$\leq n \max\{\mathrm{dia}^2(u_i)\}.$$

We also have that $(d_{min})^2 \geq \min\{\mathrm{dis}^2(u_i,u_j)\}$, where $\mathrm{dis}(u_i,u_j)$ was defined in equation (8), thus

$$\frac{\sigma_{opt}}{n * (d_{min})^2} \leq \frac{\max_i \{\mathrm{dia}^2(u_i)\}}{\min_{ij} \{\mathrm{dis}^2(u_i,u_j)\}}$$

Using equation (6) and (10), we get:

$$S \leq \frac{1}{(D_1)^2}.$$

Evidently, S becomes arbitrarily small as D_1 grows without bound. As mentioned, it has been proved by Dunn [13] that if $D_1 > 1$ the hard c-partition is unique. Thus, if the data set has a distinct substructure and the fuzzy partition algorithm has found it, then the corresponding $S < 1$.

This is consistent with equation (13). But a more interesting question is whether there exists a value such that S smaller than this value indicates the existence of a distinct substructure and the discovery of a unique fuzzy c-partition? From the above discussion, this seems to be very likely, and indeed we show below (theorem 2) that this is the case. Let us define some terms first.

Definition 9: Let fu be any local optimal fuzzy c-partition, hu being the corresponding hard c-partition. Since $S \leq 1/(D_1)^2$, there exists a value ∂_{fu} ($\partial_{fu} \leq 1$) such that $S = \partial_{fu}/(D_1)^2$. This ∂_{fu} is called the *fuzzy functional coefficient*.

Clearly, each fuzzy partition corresponds to one ∂_{fu}. We denote the minimum functional coefficient as ∂_{min}, that is, $\partial_{min} = \min_{fu} \{\partial_{fu}\}$. Thus, for all fu we have:

$$\partial_{min} \leq \partial_{fu}. \qquad (14)$$

Theorem 2: Suppose the fuzzy clustering algorithm is stable, then for every c, $2 \leq c < n$, whenever $S < \partial_{min}$, the corresponding fuzzy partition is unique. Furthermore, this partition is globally optimal.

Proof: let fu be any local optimal fuzzy c-partition with function S and hu the corresponding hard c-partition with separation index D_1. According to theorem 1 we have $S \leq 1/(D_1)^2$. And there is a ∂_{fu} such that

$$S = \partial_{fu}/(D_1)^2.$$

Thus, by equation (14),

$$S \geq \partial_{min}/(D_1)^2,$$
$$D_1 \geq (\partial_{min}/S)^{\frac{1}{2}}.$$

From the above we can explicitly see that if $S < \partial_{min}$, $D_1 > 1$. Since it has been proven that there is a unique hard c-partition if $D_1 > 1$, we can say that if the fuzzy c-partition validity function $S_{fu} < \partial_{min}$ the hard c-partition hu corresponding to fu is unique, that is,

$S_{fu} < \partial_{min} \Longrightarrow$ unique hard c-partition hu.

Now we prove that $S < \partial_{min}$ corresponds to a unique fuzzy c-partition. In fact, suppose it does not, then there exist at least two different local optimal fuzzy c-partitions fu, fv and the corresponding hard c-partitions, hu, hv such that $S_u < \partial_{min}$, $S_v < \partial_{min}$, that is,

$$S_u < \partial_{min} \Longrightarrow hu,$$
$$S_v < \partial_{min} \Longrightarrow hv.$$

But we have shown that if $S < \partial_{min}$, the corresponding hard c-partition is unique, therefore hu = hv. Since by hypothesis the fuzzy c-partition algorithm we used is convergent, for the same initial condition, it will converge to the same local optimal solution. Let us choose the hard partition results of hu and hv respectively as the initial values of the fuzzy c-partition algorithm. Since hu = hv, thus fu = fv. This means that the fuzzy c-partitions fu and fv must be the same partitions if $S < \partial_{min}$, and S_u must be the same as S_v. Thus there exists at most one fuzzy c-partition such that $S < \partial_{min}$ and this indicates its uniqueness and global optimality.

The above proof shows that ∂_{min} is very important. But unfortunately we cannot easily calculate it, although its existence is known.

V. Examples

A. Validity function S and separation index D_1

Fig. 1 depicts a set X of n = 21 two-dimensional data points arranged in c = 3 visually apparent clusters. The fuzzy c-means algorithm has been carried out using Euclidean norm, m=2.0, epsilon = 0.005 for c= 2,3,4,5, and 6. The nearest hard c-partition obtained from X in the sense of maximum membership is obtained using definition 8. Table 1 compares various results from Bezdek [2] with our calculation of S.

From table 1, we see that the hard 3-partition of X displayed in column 5 is the unique CS hard clustering of X since $D_1 > 1$. In other words, the three visually apparent clusters of fig 1 are unique CS clusters. We see that the hard partitions in table 1 suggested by the fuzzy c-means for c≠3 are not of this type. Observe also that F, H, and S identify c* =3 quite strongly, but not P which indicates c* =6. Furthermore, since $S \leq 1/(D_1)^2$ there is a ∂ such that $S = \partial * /(D_1)^2$. We use the values of S and D_1 from table 1 to calculate ∂ to see how the S and ∂ relate to each other. The values of D_1, S and ∂ are listed in table 2.

To sum up, we have seen (table 1, column 5) that there exist a unique hard c-partition when c* =3. We also see (table 2) that there exists a ∂_{min} such that $S < \partial_{min}$ indicates a unique fuzzy c-partition, that is a global optimal solution. In fact from table 2, $S < \partial$ appears for c* =3 only, that is, S = 0.02 < 0.09. This is consistent with theorem 2 and quite plausible since if the data really have a distinct substructure, the fuzzy algorithm should yield relatively hard results too. Thus we can say that c* = 3 is a unique fuzzy c-partition identified by S as rigorously as c*=3 is a unique hard c-partition identified by D_1. We also give an example to compare the results obtained from F and S [24].

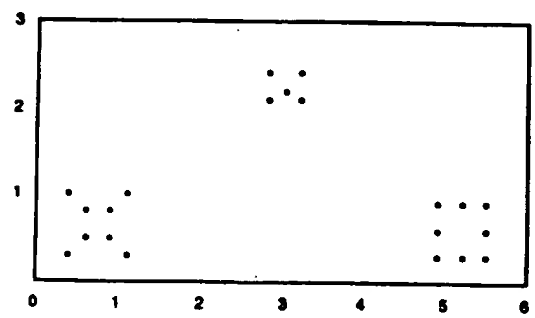

Fig 1. Data set for example A

Table 1

num of clusters C	partition coefficient F	partition entropy H^	proportion exponent P^^	CS function S	CS index D_1	CWS index D_2
2	0.85	0.23	62.3	0.06	0.05	0.03
3	0.97*	0.09*	179.4	0.02*	2.17*	2.17*
4	0.85	0.26	225.0	1.08	0.33	0.33
5	0.75	0.43	207.0	0.21	0.33	0.33
6	0.72	0.50	278.9*	0.45	0.33	0.31

Notes: ^ for definition refer to [2]; ^^ refer to [18];
* indicates the 'best' partition.

Table 2

num of clusters C	CS function S	CS index D_1	∂	> S<∂
2 0.06	0.05	0.0002	>	
3 0.02*	2.17*	0.09	<	
4 1.08	0.33	0.10	>	
5 0.21	0.33	0.02	>	
6 0.45	0.33	0.05	>	

Notes: * indicates the 'best' partition.

B. Application To Computer Color Vision

Cluster analysis has been playing an important role in solving many problems in pattern recognition and image processing. For example, it is used for feature selection in Jain and Dubes and for image segmentation for range image in Hoffman and Jain [26]. Image segmentation is a very critical step in image processing because errors at this stage influence feature extraction, classification, and interpretation at later stages.

In this section, we describe an application of our clustering criterion to color image segmentation for recognition of defects in integrated circuit (IC) wafers. The features of IC wafers are inherently colorful because of the interference effects taking place on the thin films which make up the IC structures [27]. Certain classes of IC defects can be detected by the use of colors which are otherwise not possible to detect in grey-scale image processing [27]. Various IC patterns manifest different colors due to the varying thicknesses in their structure.

In particular, we are interested in color ring defect recognition. A color ring defect is formed by a particle on the IC wafer causing a nonuniform thin film thickness surrounding the particle. The interference of different light wave lengths forms several co-centered color rings. The maximum number of rings among the colors reflects the size of the defect. Our task is to segment the color ring defect image and find the number of colors in the image and number of rings strongly formed in each color.

For the color ring defect problem, Barth [27] used a clustering method together with an appropriate distance threshold to successfully detect some defects. However, the performance of the method depends very much on the choice of the distance threshold. An improper choice of the threshold may lead to erroneous partitions. Furthermore, the right choice of such a threshold is unknown a priori. Another concern is the potential of the method to give a very large number of partitions which may not be very useful for solving problems such as the "ring" defect problem.

Here, we describe the use of our clustering validity criterion to the segmentation of two color ring defect images taken from real samples. The image is of 512X480 pixels. Since each defect occupies a very small part of the image, a focusing_of_attention strategy [27] is employed, that is we only segment a small part of the image which contains one color ring defect. In such a way, the size of data to be processed can be largely reduced and computing time saved. Notice that there are some noises in both images. To reduce the noise effect, a threshold for pixel density in color space is used. Only those colors with density larger than the threshold are processed by using the clustering algorithm. The choice of this threshold does not essentially affects the results. The remaining data points are assigned to the nearest cluster. After segmentation, all pixels in each cluster are assigned the color value of the centroid of that cluster.

Fig. 2 is the picture of a color ring image which contains two color ring defects. We focus our attention on the upper left color ring inside the window shown in Fig.2. There are two main visually apparent colors ("red", "green") in the window. This part of image needs to be segmented in order to find the number of colors existing in this defect and number of color rings formed in each color. The problem is to find the number of clusters in the color space that best partitions or segments the image. From the image inside the window, 290 distinct data points in color space are obtained by using a pixel density threshold of 4.

The partitioning result is shown in Fig.3. Our validity criterion identifies the 2-partition as the best solution, thus correctly reflecting the two visually apparent colors in the image. Fig. 4 (a-d) shows the results of 2 to 6 partitions with each segmented color displayed in a separate window. In Fig. 4a, the lower left window shows the segmented image for the window above it, and the colors in the right two windows are "red" and "green", respectively, which are the actual apparent color structure in the image. Notice that edges separating the two colors are very clear and the rings formed in each color is quite strong. There are three clear rings in the red color window and three clear rings in the green color window. Thus, the maximum number of rings among the two colors is three and the total number of rings in this defect is six.

The second best solution (see Fig.3) is c=3. Each segmented color for 3-partitions is shown in a separate window in Fig.4b. A third color (dark brown) is split from the 2-partitions. However, one can see the existence of noises. The number of rings in this color is not clear at all and the edges are not smooth either. Similarly (or even worse), all the other c-partitions with c>3 produce irregular segmentations or weaker rings.

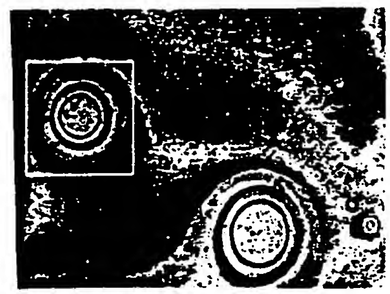

Fig.2 Color ring image for example B

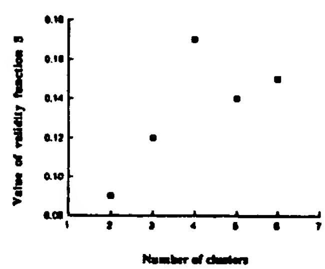

Fig.3 Partition result for example B

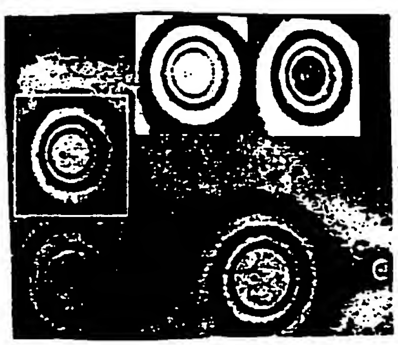

Fig.4a 2-partitions

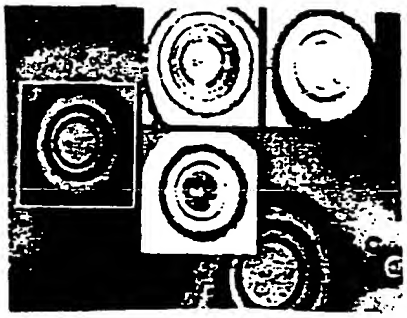

Fig.4b 3-partitions

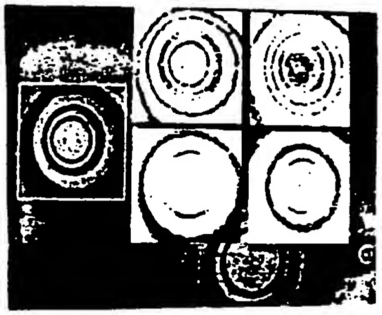

Fig.4c 4-partitions

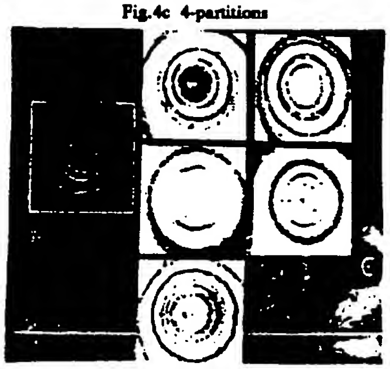

Fig.4d 5-partitions

VI.. Conclusion

In summary, the main result of the paper is the introduction of a new strategy for the determination of the degree of validity of a fuzzy partition. The strategy has been mathematically justified via its relation to hard partition validity measures. We have derived the relationship (theorem 1) between this fuzzy validity function and the most general, and well defined, hard clustering validity function ('separation index' of Dunn for which the condition of uniqueness has already been established). By using this relationship, we have also proven (theorem 2) the existence of a unique fuzzy c-partition produced by the fuzzy validity function. Examples of applications to segmentation of color image for IC defects give encouraging results. The main advantage of the new strategy to determine validity is its computability which allows applications to 'real-time' engineering systems, such as color vision system, robotic systems, and distributed perception networks. These applications are currently under investigation [6].

Acknowledgements

The authors wish to thank Degang Chen for many valuable discussions and suggestions, and Bruce Anderson and professor Susan Hackwood for making the color vision system available. The research was supported in part by NSF grant 08421415.

References

[1] A. K. Jain, *Handbook of pattern recognition and image processing*, Chapter 2, Academic Press, New York, 1988.

[2] J. C. Bezdek , *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York and London, 1981.

[3] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice hall, New Jersey, 1988.

[4] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering", *IEEE Transactions on PAMI*, Vol. 11. No 7. July 1989.

[5] I. Gath and A. B. Geva, "Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions," *Pattern Recognition letter* 9, pp. 77-86, N.-Holland, 1989.

[6] X. Xie and G. Beni, "Distributed hierarchical decision fusion with cluster validity," *Proc. of the IEEE International Conf. on Systems, Man and Cybernetics*, Los Angeles, pp.689, Nov. 1990.

[7] A. M. AL-Bassiouni, "Optimal signal processing in distributed sensor systems," *Ph.D. Thesis*, Naval Postgraduate School, Monterey, California, 1987.

[8] A. Kandel, *Fuzzy techniques in pattern recognition*, John Wiley & Sons, New York, 1982.

[9] E. Backer and A. K. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Transactions on PAMI*, Vol. PAMI-3. No 1. January,1981.

[10] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, John Wiley, New York, 1973.

[11] N. E. Day, "Estimating the components of a mixture of normal distributions," *Biometrika* 56: 463-474, 1969.

[12] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," *Multivar. Behav. Res.* 5, 329-350. 1970.

[13] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions", *J. Cybern.*, Vol. 4-1, pp. 95-104, 1974.

[14] D. L. Davies and D. W. Bouldin, " A cluster seperation measure", *IEEE Transactions on PAMI*,Vol. PAM-1. No 2. April 1979.

[15] A. K. Jain and J. V. Moreau, " Bootstrap technique in cluster analysis," *Pattern Recognition*, Vol. 20, No. 5, pp. 547-568, 1987.

[16] J. C. Bezdek, "Numerical taxonomy with fuzzy sets", *J. Math. Biol.*, Vol. 1, pp. 57-71, 1974.

[17] J. B. Bezdek, "Mathematical models for systematics and taxonomy," *in Proc. 8th Int. Conf. on Numerical Taxonomy*, G. Estabrook, Ed. Freeman, San Francisco, CA, pp. 143-166, 1975.

[18] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets Syst.*, vol. , pp. 177-185,1981.

[19] M. P. Windham, "Cluster validity for fuzzy c-Means clustering algorithm," *IEEE Transactions on PAMI*,Vol. PAMI-4. No 4. July,1982.

[20] G. Libert and M. Roubens "New experimental results in cluster validity of fuzzy clustering algorithms" in J. Janssen, J-F. Macrotorchino and J-M. Proth (eds) *New Trends in Data Analysis and Applications*, pp. 205-218, North Holland, 1983.

[21] M. P. Windham, H. Bock, H. F. Walker, "Clustering information from convergence rate," *2nd Conf. International Federation Classification Soc.* pp. 143, Washington, 1989.

[22] R. Gunderson, "Application of fuzzy ISODATA algorithms to star tracker pointing systems", *Proc. 7th Triennial World IFAC Congress*, Helsinki, pp. 1319 -1323, 1978.

[23] J. C. Dunn, "Indices of partition fuzziness and detection of clusters in large data sets," *Fuzzy Automata and Decision Processes*, Elsevier, New York, 1977.

[24] A. K. Jain and R. Dubes, "Feature definition in pattern recognition with small sample size," *Pattern Recognition*, Vol. 10, pp. 85-97, 1978.

[25] X. Xie and G. Beni, "A new validity criterion for fuzzy partitions,"*IEEE Transactions on PAMI*, June 1991.

[26] R. Hoffman and A. K. Jain, "Segmentation and Classification of range images," *IEEE Transactions* [27] M. Barth, "Robotic attentive sensing and its application to integrated circuit wafer inspection", *Ph.D. dissertation*, University of California, Santa Barbara, Aug. 1989.